

A Comparison of the Frequency of Number/Punctuation and Number/Letter Combinations in Literary and Technical Materials

Abstract

For the past 12 years, a research project has been underway to study the development of a unified braille code for the English language, i.e., a single code for all literary, mathematics, scientific, computer, and technical materials. One of the major decisions that had to be made early in the Unified English Braille Code (UEBC) Research Project was the position of numbers in the braille cell. The decision was made to use upper numbers, as used now in the literary code and used throughout the world. One of the reasons for making this decision was the assumption that numbers are followed more frequently by punctuation than by letters in all types of materials, which would save space when transcribing materials into the UEBC. This study sought to verify this assumption by generating a computerized count of all the number/letter and number/punctuation sequences in selected general and technical texts.

The results corroborated the assumption that a sequence of number/punctuation occurs much more frequently than that of a number/letter sequence. In the present mathematics code used in the United States, Canada, and New Zealand (which uses lower numbers), the punctuation indicator would be required 97.81 times more frequently than the letter indicator would be required in UEBC.

A COMPARISON OF THE FREQUENCY OF NUMBER/PUNCTUATION AND NUMBER/LETTER COMBINATIONS IN LITERARY AND TECHNICAL MATERIALS

Submitted by Darleen Bogart, Frances Mary D'Andrea, Alan Koenig

The Unified English Braille Code (UEBC) project is an international research effort, initiated by the Braille Authority of North America (BANA) in 1991 and continued since 1993 by the International Council on English Braille (ICEB). The participating countries are Australia, Canada, New Zealand, Nigeria, South Africa, the United Kingdom, and the United States, all of which have braille experts working as volunteers on the development of UEBC.

A number of committees have worked on various aspects of the UEBC. The main work for the development of the code has been the responsibility of Committee II – Extension of the Base Code. The four US members who began deliberations on the BANA project became members of the international committee, which was expanded to include one additional member from every participating country. Five of the original eight members were braille consumers, while seven of the current eight-member group are braille consumers. In fact, all of the four other working committees have always had a majority of braille consumers as members. Also represented on each committee are educators who teach braille and braille producers/transcribers. The same basic principles that were articulated at the beginning of the UEBC project were accepted when it was internationalized by ICEB. Literary braille, which is similar in all the participating countries, was to be used as the base code, which was to be extended in such a way that all the technical symbols could be imbedded in it and be unambiguous. (Music is not part of the project, as it is already an international braille code.)

One important aspect of the proposed UEBC was the decision made early in the code development process to use literary numbers that occupy the upper part of the braille cell (dots 1, 2, 4, 5). This option was considered preferable to Nemeth numbers used for mathematics and scientific materials, that occupy the lower part of the braille cell (dots 2, 3, 5, 6), or French (dot 6) numbers which use the dot configurations of the upper numbers 1-9 with the addition of dot 6 within each cell to create the digits 1-9, with another configuration for 0, as used in some European computer codes. A question asked frequently, especially in countries using the BANA codes, is why upper numbers were chosen as the basis of the number system in UEBC. The merits of each of the three numbering systems were thoroughly investigated and discussed by the original BANA committee that worked on developing a “base code” and also by the international committee after it took charge of the project.

The appeal of “dot 6” numbers was their complete unambiguity-- a number sign was not needed with them. Braille punctuation indicators and letter signs were not needed. The

number of cells required to represent numbers, and number/letter, number/punctuation would be the lowest of the three options. The main factor against their acceptance was that the loss of up to ten contractions (e.g., ch, gh, sh, th, wh, ed, er, ou, ow), was felt to be too great a change to English Braille American Edition (EBAE).

The appeal cited for “lower” numbers was that fewer indicators would be required in mathematics when numbers and letters were not in the same part of the braille cell (i.e., a letter could follow a number without the need for a letter sign). Lower numbers are the same dot configuration as many of the punctuation signs (i.e., comma is 1, semi-colon is 2, colon is 3, period is 4, exclamation mark is 6, parenthesis is 7, question mark is 8). Thus the use of lower numbers next to punctuation marks that conflict with numbers would require the use of a punctuation indicator before the punctuation mark. Small samples of mathematics and technical materials analyzed in those early years of the project indicated that numbers generally come in contact with punctuation more often than with alphabet letters a-j.

The case for retaining the upper numbers of the base code, literary braille, was further supported by the fact that upper numbers are international and are understood in all countries. Therefore, it would be less “disruptive” (i.e. would require fewer new symbols for readers) if the proposed code used literary, upper cell numbers rather than requiring lower numbers to be used in all types of materials. However, a much larger sample needed to be analyzed from a wide variety of texts to more fully test this hypothesis and to determine whether number/letter combinations or number/punctuation combinations occur more frequently in technical and student texts and general interest materials.

Method and Procedures

The Canadian National Institute for the Blind (CNIB), through its Library for the Blind, undertook this project in the summer of 2002. Mr. Tom Keith, who has considerable programming experience and is also a volunteer braille transcriber certified in both the literary and Nemeth codes, developed a computer program that counted the occurrences of number/letter and number/punctuation combinations in the regular, superscript, or subscript positions in 16 textbooks for a total of 8,429 pages of text (see Chart 1). The sample included 4,556 pages of technical material that required the Nemeth code and 3,873 pages of other texts that did not require the Nemeth Code, i.e., used the literary code. Titles were selected that had a considerable number of numerals. Each of the texts chosen had copyright dates no older than 10 years.

The program was designed to identify series of dots including numbers followed by punctuation and by letters in the brf versions of these titles. This was relatively simple for the material brailled in literary code as each number series required a number sign. The material brailled in the Nemeth Code presented more difficulty as not every series which includes a number begins with a number sign. Even though page numbers appeared on every braille page in the brf file, they were not included in the count.

The program recognized all number/letter combinations and all number/punctuation combinations. Totals were given for each category within each code type (literary and Nemeth) as well as the accumulated totals for each type of material. The program arranged them in order of frequency. The letters occurring immediately following numbers in these texts were: a-j, A-J, k-z, K-Z, dg, nd, rd, st, th, sin and cos. The punctuation marks occurring immediately following numbers were: period, comma, semicolon, colon, question mark, exclamation mark, closing parenthesis, closing brackets, closing angle bracket, closing curly brace, hyphen, dash, percent sign, degree symbol, apostrophe s. The internal commas and decimal points were not counted.

Results

The technical material had 79,827 numbers with 10 or more instances of each combination. The literary texts had 25,223 numbers with three or more instances of each combination. If the number of combinations is cut off at 10 for this group of literary texts so that it matches the technical texts, the total number of instances would be reduced by 310 to 24,913. Thus the frequency counts used in this study reflected this lower number, i.e. only those with 10 or more instances of each combination were considered.

The number/letter combinations a-j used in this study would require a letter sign before the letter when following an upper number, as they are in current literary code rules and as they would be in UEBC. For example, the phrase 4d would require a letter sign after the number 4 to give letter meaning to the d. All other lower case letters (k-z) require the letter sign in current literary code but, because they are not in conflict with numbers, they do not require the letter sign in UEBC. All upper case letters (A-Z) in the present literary code require a letter sign but, because they are preceded by the braille capital sign, they are not in conflict with numbers and would not require a letter sign in UEBC.

The total incidence in these samples of lower case letters a-j following a number was 212. One hundred and eighty-two of these were in technical material, which do not require a letter sign when used with “lower” numbers. But in UEBC, all 212 letters a-j would require a letter sign following numbers which are all in the upper part of the braille cell (see Table 1).

The number/punctuation combinations counted are those that would require a punctuation indicator before the punctuation when lower numbers are used, as they would be in the Nemeth Code. For example, the phrase 4. in Nemeth Code would require a punctuation indicator after the number 4 to give punctuation meaning to the period. The literary punctuation marks retained in the Nemeth Code, with the same dot configurations as lower numbers, that require the punctuation indicator are: period, semicolon, colon, question mark, exclamation mark and apostrophe s. (The literary comma is not used in the Nemeth code and has been replaced by dot 6, which does not have a number meaning and thus does not require a punctuation indicator before it. The Nemeth Code has other punctuation marks not available in the literary code, which do not conflict with lower numbers, e.g. closing parenthesis, bracket, angle bracket and curly brace.) A basic principle of UEBC is that each print symbol will have one braille representation. As a

result some punctuation marks have been changed rather than initiating the use of a punctuation indicator (e.g., opening and closing enclosure symbols, such as parenthesis, brackets, braces).

The total number in these samples of affected punctuation marks following a number was 36,505. Of these, 15,770 were in literary code material, which do not require a punctuation indicator when used with upper numbers. That left 20,735 instances, the total count for technical materials that currently use lower numbers requiring a punctuation indicator before the punctuation mark that follows a number. But in UEBC no punctuation indicators would be required to show the difference between numbers and punctuation marks because upper numbers are used exclusively (see Table 2).

If the decision of the UEBC base committee had been to adopt lower numbers in building the code, then there would be no need for letter signs in the number-letter combinations a-j, but there would have been the need for 30,509 punctuation indicators for number-letter combination. If the base code committee had adopted the Dot-6 number system, neither letter signs nor punctuation indicators would have been required. But as stated earlier, the Dot-6 numbers are less familiar for most braille users in English-speaking countries. Table 3 shows the numbers of indicators that would be needed for a lower number and a Dot-6 number systems.

Conclusion

The results indicate that a much higher instance of number/punctuation combinations than number/letter combinations were found in all texts. Thus a far greater number of punctuation indicators (20,735) were required when lower numbers are used than letter signs (499) when upper numbers are used. When UEBC rules are applied, the number of letter signs needed drops to 212, and no punctuation indicators are required. The incidence of punctuation indicators required in the present BANA technical code (Nemeth code) occurs 97.81 times as frequently as the letter sign is required in UEBC. That count reinforces the previous count results and gives strong support for the decision to use upper numbers in the UEBC.

CHART 1
Texts Used in Frequency Count

| Titles of Texts | Print Pages |
|--|--------------------|
| Literary Code | |
| Civics: Participating in a Democratic Society | 205 |
| Introduction to Economics: A Canadian Analysis | 726 |
| Insights: Succeeding in the Information Age | 312 |
| Marketing Dynamics | 412 |
| Gage Physical Geography 7: Discovering Global Systems and Patterns | 282 |
| Psychology, Canadian Edition | 728 |
| Regional Dynamics: A Geography of Travel & Tourism | 255 |
| Science Everywhere 3 | 247 |
| Science Everywhere 4 | 254 |
| Scienceplus 9 | 452 |
| Total | 3,873 |
| Nemeth Code | |
| Probability and Random Processes for Electrical Engineering; Second Edition | |
| Quest 2000: Exploring Mathematics, Grade 5 | 286 |
| Quest 2000: Exploring Mathematics, Grade 6 (Revised Edition) | 271 |
| Foundations of Mathematics 10 | 488 |
| Nelson Chemistry (British Columbia Edition) | 1,792 |
| Calculus: Early Transcendentals (Fifth Edition) | 1,123 |
| Total | 4,556 |
| TOTAL PAGES | 8,429 |

Table 1
Frequency of Number/Letter Combinations

| | Occurrences | | Letter sign required in | |
|--------------|-------------|--------------|-------------------------|------------|
| | Literary | Nemeth | Present BANA Codes | UEBC |
| a-j | 30 | 182 | 30 | 212 |
| A-J | | 35 | | |
| k-z | 400 | 2,030 | 400 | |
| K-Z | | | | |
| dg | 69 | | 69 | |
| Total | 499 | 2,247 | 499 | 212 |

Table 2
Number/Punctuation Frequency

| | Occurrences | | Punctuation indicator required in | |
|-----------------|---------------|---------------|-----------------------------------|----------|
| | Literary | Nemeth | Present BANA codes | UEBC |
| . period | 8,302 | 19,935 | 19,935 | |
| , comma | 5,996 | 12,915 | | |
| ; semi-colon | 1,023 | 301 | 301 | |
| : colon | 380 | 280 | 280 | |
| ? question mark | 69 | 219 | 219 | |
| Total | 15,770 | 33,650 | 20,735 | 0 |

Table 3
Number of Letter Signs and Punctuation Indicators Required
in UEBC if Lower Numbers and Dot 6 Numbers Were Used

| | Letter Signs Required in UEBC were Adopted with | | Punctuation Indicators Required in UEBC Adopted with | | |
|--------------|---|------------------|---|------------------|------------------|
| | Lower Numbers | Dot 6 Numbers | | Lower Numbers | Dot 6 Numbers |
| a-j | | | Period | 28,237 | |
| A-J | | | Comma | 5,996 | |
| k-z | | | Semicolon | 1,324 | |
| K-Z | | | Colon | 660 | |
| | | | Question mark | 288 | |
| Total | 0 | 0 | | 36,505 | 0 |